Capturing Spatiotemporal Signaling Patterns in Cellular Data with Geometric Scattering Trajectory Homology

Dhananjay Bhaskar[1†], **Jessica Moore**[1†], **Feng Gao**[2], **Bastian Rieck**[3], **Firas Khasawneh**[4], **Elizabeth Munch**[5,6], **Valentina Greco**[1,7,8,9,10*], and **Smita Krishnaswamy**[1,10,11,12,13*]

[1]Department of Genetics, Yale School of Medicine
[2]Mailman School of Public Health, Columbia University
[3]Institute of AI for Health, Helmholtz Munich, Technical University of Munich
[4]Department of Mechanical Engineering, Michigan State University
[5]Department of Computational Mathematics, Science and Engineering, Michigan State University
[6]Department of Mathematics, Michigan State University
[7]Department of Cell Biology, Yale School of Medicine
[8]Department of Dermatology, Yale School of Medicine
[9]Yale Stem Cell Center, Yale University
[10]Yale Cancer Center, Yale University
[11]Department of Computer Science, Yale University
[12]Program for Computational Biology and Bioinformatics, Yale University
[13]Program for Applied Mathematics, Yale University

[*]Corresponding authors: `valentina.greco@yale.edu, smita.krishnaswamy@yale.edu`; [†] Co-first authors

March 22, 2023

Abstract

Cells communicate with one another through a variety of signaling mechanisms. Exchange of information via these mechanisms allows cells to coordinate their behaviour and respond to environmental stress and other stimuli. To facilitate quantitative understanding of complex spatiotemporal signaling activity, we developed *Geometric Scattering Trajectory Homology*, a general framework that encapsulates time-lapse signals on a cell adjacency graph in a low-dimensional trajectory. We tested this framework using computational models of collective oscillations and calcium signaling in the *Drosophila* wing imaginal disc, as well as experimental data, including *in vitro* ERK signaling in human mammary epithelial cells and *in vivo* calcium signaling from the mouse epidermis and visual cortex. We found that the geometry and topology of the trajectory are related to the degree of synchrony (over space and time), intensity, speed, and quasi-periodicity of the signaling pattern. We recovered model parameters and experimental conditions by training neural networks on trajectory data, showing that our approach preserves information that characterizes various cell types,

tissues and drug treatments. We envisage the applicability of our framework in various biological contexts to generate new insights into cell communication.

1 Introduction

Cell signaling plays a critical role in orchestrating the complex interactions and processes necessary for the proper functioning and survival of organisms. This intricate communication system allows cells to perceive and respond to their ever-changing environment by transmitting and processing information through a series of biochemical reactions. Cells regulate the timing and location of signaling activity to coordinate a myriad of cellular functions, including embryonic development, tissue repair, and immune responses [1–5]. By fine-tuning the spatial and temporal characteristics of signaling, cells are able to communicate effectively, adapt to changing environments, and maintain homeostasis. Impairment of signaling mechanisms is a hallmark of various diseases, including cancer, developmental disorders, and degenerative diseases.

Although many signaling molecules and pathways have been identified, the dynamics of signaling processes, including their initiation, propagation, termination, and adaptation, are not yet fully understood. Cellular signaling networks are highly complex, involving numerous molecules, pathways, and feedback loops. This complexity makes it challenging to study individual components and their interactions within the system. Moreover, these networks often show emergent properties that cannot be easily determined from the behavior of their individual components. Signaling dynamics are often non-linear and context-dependent, making it difficult to predict their outcomes based on simple linear models or assumptions [6]. The outcome of a signaling event may depend on the cellular context, including the presence of other signaling molecules, the cellular environment, and the history of previous signaling events. The spatial distribution and temporal dynamics of signaling molecules can be highly heterogeneous within a tissue. Heterogeneity can arise from various factors, such as differences in the expression and localization of signaling molecules in neighboring cells, as well as variations due to noise, changes in the extracellular environment or an external stimulus. This heterogeneity adds another layer of complexity to the study of signaling dynamics.

The complexities and current lack of understanding in signaling dynamics necessitates the development of advanced computational methods for the quantitative analysis of spatiotemporal signaling. Incorporating cell-cell connectivity at multiple scales and the temporal dynamics of the signal is crucial for a comprehensive understanding of these intricate processes. By accounting for the multi-scale nature of cellular interactions and the temporal fluctuations of signaling events, we can elucidate the underlying mechanisms that govern cellular communication and coordination. This, in turn, can shed light on various pathological conditions and inform the development of targeted therapeutic strategies. Furthermore, a robust computational framework capable of integrating multi-scale connectivity and temporal dynamics would enable researchers to generate predictive models, fostering a more efficient and effective approach to understanding the functional significance of various signaling pathways in diverse cellular contexts.

In this paper, we develop a framework for analyzing spatiotemporal patterns in cell signaling, called *Geometric Scattering Trajectory Homology* (GSTH), that integrates geometric scattering and topological data analysis (TDA) to provide a comprehensive understanding of complex cellular interactions. This combination allows for the effective capture of both local and global patterns, as well as a robust analysis of the underlying topology. First we construct a graph from cell positions with the time-lapse signal as node features. We then apply geometric scattering to quantify the spatiotemporal features of signaling dynamics. Geometric scattering employs wavelet-based transformations to extract multiscale representations of the signaling data, capturing the intricate hierarchical structures present in the spatial organization of cells and the temporal evolution of signaling events [7, 8]. These features provide a rich, multiscale characterization of the spatiotemporal patterns that are essential for understanding the complex dynamics of cellular processes.

We obtain a low dimensional trajectory of the signaling dynamics using PHATE [9], a manifold-geometry preserving dimensionality reduction technique. The trajectory is generated by embedding geometric scattering features for each time point. Next, we use TDA to study the higher-order properties of the PHATE trajectory by computing the persistent homology of the signal embeddings. TDA is a powerful tool for investigating the shape and connectivity of data, providing a robust, scale-invariant representation of the underlying structure. Persistent homology, a key technique within TDA, is used to capture the topological features of the data across multiple scales, revealing hidden patterns and relationships in cell signaling dynamics.

We test the proposed framework on simulations of collective oscillatory behavior and calcium signaling in the *Drosophila* wing imaginal disc. First we simulate a system of coupled oscillators at varying coupling strengths to generate signals with different degrees of synchrony. We show that the signal embeddings produced by GSTH converge over time as the oscillators get synchronized, resulting in shorter, denser trajectories characterized by Betti numbers in the persistent homology. We show that PHATE trajectories preserve spatiotemporal dynamics by predicting the coupling strength from the PHATE trajectories and topological features. Next we simulate four different patterns of calcium signaling by varying the rate of production of inositol trisphosphate (IP_3) in a mathematical model of intercellular calcium signaling in the *Drosophila* wing imaginal disc. We show that intermittent spiking and uncoordinated transient signaling produce gaps in the PHATE trajectory, which are detected using TDA. On the other hand, spatially and temporally coordinated signaling produces smooth trajectories without gaps. We demonstrate that the model parameters used to generate the simulation can be recovered using a neural network trained on geometric scattering features, highlighting the predictive capability of our model.

Finally we apply GSTH to three experimental datasets, consisting of: (1) calcium signaling in the basal stem cell layer of the mouse epidermis, (2) ERK signaling in oncogenic mammary epithelial cells, and (3) calcium signaling in the prefrontal visual cortex of mice. We show that GSTH can accurately classify control versus perturbations in the first two datasets, and visual stimuli presented to mice in the third dataset.

Geometrical and topological analysis of spatiotemporal signaling can be used to understand the complex dynamics of biological systems, such as how cells process information and coordinate their behavior. By combining geometric scattering and topological data analysis, our framework is able to integrate local and global information, as well as geometrical and topological properties, to provide a more comprehensive understanding of spatiotemporal patterns in cell signaling. This innovative approach paves the way for the development of predictive models, improved therapeutic strategies, and a deeper understanding of the complex processes governing cellular communication and coordination.

2 Results

2.1 The GSTH architecture

GSTH comprises of three steps: (1) spatial embedding of the signal using geometric scattering, (2) temporal embedding of the multiscale geometric features using PHATE to obtain low-dimensional trajectories, and (3) computation of topological features of the trajectory using persistent homology. Next, we describe these steps in detail.

2.1.1 Spatial embedding

First, we encode spatial variations in the signaling pattern at a given timepoint, t_i, using geometric scattering, an unsupervised method for generating embeddings for graph-structured data [7]. This involves multiscale diffusion of the signal, $X(t_i)$, by performing random walks on a cell neighborhood graph, $G(V, E)$, constructed from cell position data. The graph consists of vertices, $V = \{v_1, v_2, \cdots, v_M\}$, represent individual cells and edges connecting cells that lie within some predefined neighborhood distance. We represent the graph using the adjacency matrix, $\mathbf{A}$, and the degree matrix, $\mathbf{D}$. We use the notation $x(v_l, t_i)$ to denote the signaling activity at cell v_l at time t_i.

The geometric scattering transform applies a cascade of graph wavelet transforms followed by a nonlinear modulus operation [7, 8]. The design of graph wavelets is based on the diffusion operator constructed from lazy random walks, $\mathbf{R} = \frac{1}{2}\left(\mathbf{I} + \mathbf{A}\mathbf{D}^{-1}\right)$, over a graph, i.e.,

$$\mathbf{\Psi}_0 = \mathbf{I} - \mathbf{R}\,, \quad \mathbf{\Psi}_j = \mathbf{R}^{2^{j-1}} - \mathbf{R}^{2^j} = \mathbf{R}^{2^{j-1}}(\mathbf{I} - \mathbf{R}^{2^{j-1}})\,, \quad j \geq 1\,. \tag{1}$$

The multi-scale nature of graph wavelets allows the geometric scattering transform to traverse the entire graph in one layer, which provides both local and global graph features. Summation of the signal responses is used to obtain invariant graph-level features. Since the summation operation could suppress the high frequency information, it could be complemented by using higher order summary statistics of signal x. Due to the iteration of applying graph wavelets followed by a nonlinear modulus operation, geometric scattering transforms can be constructed as in a multi layer (or multi order) architecture. Specifically, the zeroth-order scattering coefficients are calculated by taking statistical moments of the summation of signals, and the first order features are obtained by applying a graph wavelet, which aggregates multiscale information of the graph. Second-order geometric scattering features can further augment first order features by iterating the graph wavelet and absolute value transforms. The collection of graph scattering features provides a rich set of multiscale invariants of the graph G and can be used under both supervised and unsupervised settings for graph embedding.

For a signal $X(t_i) = [x(v_1, t_i), x(v_2, t_i), \ldots, x(v_M, t_i)]$ we compute the zeroth-order scattering coefficients for each vertex/cell for timepoint t_i as follows:

$$S_0(X(v_\ell, t_i)) = \mathbf{R}^{2^J} x(v_\ell, t_i). \tag{2}$$

The diffusion operator (lazy random walks) $\mathbf{R}$ here works as a low pass filter that performs local averaging of signal in neighboring cells[7, 10]. Unlike the summation operator that averages all vertex information and suppresses the high frequency information and hence has to be retrieved by higher order statistical moments, this retains finer description of cell/vertex embeddings. Then, by concatenating the wavelet coefficients for each cell/vertex at timepoint t_i, we can obtain the corresponding timepoint embedding $S_0(X(t_i))$ for timepoint t_i. Finally, the timepoint embedding for N timepoints can be calculated and the resulting $S_0(X(t)) = \{S_0(X(t_0)), S_0(X(t_1)), \ldots, S_0(X(t_n))\}$ is a feature matrix of dimension $N \times M$, where N is the number of timepoints and M is the number of cells. We hence obtain the zeroth-order scattering coefficients for the N timepoints. The scattering transform here is a result of local averaging of wavelet coefficients.

The zeroth-order scattering features can be augmented by first-order scattering features by applying graph wavelets and extracting finer description of high frequency response of a signal $X(t_i)$ [7]. Specifically, the first-order scattering coefficients for each time point at each vertex/cell are calculated as:

$$S_1(x(j, v_\ell, t_i)) = \mathbf{R}^{2^J} |\mathbf{\Psi}_j x(v_\ell, t_i)|,\ 1 \leq j \leq J, \tag{3}$$

The value $\mathbf{\Psi}_j x(v_\ell, t_i)$ aggregates the signal information $\mathbf{x}(v_m, t_i)$ from the vertices v_m that are within 2^j steps of v_ℓ. It responds to sharp transitions or oscillations of the signal $\mathbf{x}$ within the neighborhood of v_ℓ

with radius 2^j (in terms of the graph path distance). By concatenating all the vertex/cell embeddings, we can obtain the first order scattering coefficients $S_1(X(t_i))$ for timepoint t_i.

Finally, the second-order scattering coefficients can be obtained by further applying graph wavelets and extract even finer description of high frequency response of the signal $X(t_i)$:

$$S_2(X(j, j', v_\ell, t_i)) = \mathbf{R}^{2^J} |\mathbf{\Psi}_{j'}|\mathbf{\Psi}_j x(v_\ell, t_i)||, 1 \leq j < j' \leq J \tag{4}$$

The above calculations are conducted for each timepoint and a total of N timepoints. The first-order and second-order scattering transform will generate a feature matrix of shape $N \times (M \times J)$ and $N \times (M \times \frac{J \times (J-1)}{2}))$, respectively, as timepoint embeddings for the N timepoints. Finally, the zeroth-order, first-order and second-order scattering coefficients were combined together as the embeddings for each time point $S(X(t_i))$. The scale of the wavelet J was selected based on the diameter of graphs, and the number of scattering coefficients generated depended on the graph sizes.

2.1.2 PHATE trajectories

The signal embeddings $S(X(t_i))$ from geometric scattering form a matrix of dimensions $T \times M$, where T is the number of time points in the data and M is the number of scattering coefficients for each time point. We can visualize these embeddings by applying PHATE. Following our previous description of PHATE, we calculated a distance matrix $\mathbf{D} = \|S(X(t_i)) - S(X(t_j)\|_2$ based on the Euclidean distance between time point embeddings and applied an α-decaying kernel K with a locally-adaptive bandwidth $\epsilon_{k,i}$ corresponding to the k-NN distance of the i-th data point to generate an affinity matrix $\mathbf{W}$ as well as the diffusion operator $\mathbf{P}$. The elements of $\mathbf{W}$ are given by:

$$\mathbf{W}_{i,j} = K_{k,\alpha}(i,j) = \frac{1}{2}\exp\left(-\left(\frac{\mathbf{D}_{i,j}}{\epsilon_{k,i}}\right)^\alpha\right) + \frac{1}{2}\exp\left(-\left(\frac{\mathbf{D}_{i,j}}{\epsilon_{k,j}}\right)^\alpha\right) \tag{5}$$

The decaying factor α regulates the decay rate of the kernel (smaller $\alpha \Rightarrow$ kernel with lighter tails), $\alpha = 2$ corresponding to the Gaussian. The diffusion operator $\mathbf{P}$ can then be obtained by calculating the row-sum of the affinity matrix $\mathbf{W}$ with element $\mathbf{P}_{i,j}$ giving the probability of moving from the i-th to the j-th data point in one time step. The global structure of the data can be further learned through calculating the tth power of the diffusion operator $\mathbf{P}$, which propagates affinity of the data through diffusion up to a scale of t. The optimal value t for diffusion is automatically chosen to be the knee point of the von Neumann entropy of $\mathbf{P}^t$. This diffusion operator is then log scale transformed and converted to a potential distance matrix $\mathrm{ID}(X)$ which is embedded by MDS to result in 3-D PHATE embedding coordinates $E(t) = (E_1(X(t)), E_2(X(t)), E_3(X(t)))$ for each time point t, and point cloud $E = \{E(t_1), E(t_2), \ldots E(t_n)\}$.

The 3D coordinates enable visualization of the trajectory, which reflects the time-varying patterns of Ca^{2+} fluorescence data. Thus neighbors in the PHATE embedded trajectories indicate similar signaling patterns even if they occur at distal timepoints. In fact, many of the dynamics we notice have loop-like structure, which motivates the computation of topological features described in the next section.

2.1.3 Persistent homology analysis of PHATE trajectories

To obtain a topologically invariant characterization of a given PHATE trajectory, denoted E, we calculated its persistent homology. Specifically, we calculated the persistent homology of E via a Vietoris–Rips filtration

$\mathrm{VR}_s(E)$. The Vietoris–Rips complex of E is defined as the filtered simplicial complex that contains a subset of E as a simplex if and only if all pairwise distances in the subset are less than or equal to s, i.e., $\mathrm{VR}_s(E) = \{\{n_0, ..., n_m\} \mid \forall i, j\ d(i, j) \leq s\}$. We noted here that we could also use the potential distance ID from PHATE, however we directly used the PHATE coordinates and the Euclidean distance for simplicity.

As described above, from $\mathrm{VR}_s(E)$, we obtain a set of persistence diagrams Q consisting of birth-death-dimension triples $[b, d, q]$ that describe multiscale topological features of E. Each such point corresponds to a topological feature in the trajectory, which appears at a certain birth time and disappears at a death time. Note that the times are supposed to be understood with respect to the parameter s from above. A point's distance from the diagonal therefore represents the prominence or the eponymous *persistence* of the associated topological feature; higher values indicate that the feature occurs over a large scale, thus increasing its significance. We further calculated the associated Betti curves for each Q, resulting in a simple summary curve $B(Q, q)$ for the qth dimension consisting of the number of points (b_i, d_i) in Q such that $b_i \leq s < d_i$. The Betti curve characterizes the connectivity of $VR_s(E)$ and, by extension, of the Ca^{2+} fluorescence data.

In addition, we computed pairwise Wasserstein distances between persistence diagrams for unsupervised classification of various experimental conditions. The Wasserstein distance between two persistence diagrams Q_1 and Q_2 is calculated as follows:

$$W_m(Q_1, Q_2) = \Big(\inf_{\eta: Q_1 \to Q_2} \sum_{q_1 \in Q_1} ||q_1 - \eta(q_1)||_\infty^m \Big)^{\frac{1}{m}} \tag{6}$$

where η is a bijection from Q_1 to Q_2. Specifically, we consider the diagonal Δ of persistence diagrams to have infinite multiplicity, i.e. points can be matched to the diagonal. We used Eirene to compute Wasserstein distances, which utilizes the Hungarian algorithm to find the optimal matching. It has been shown that persistence diagrams are stable [11]. Hence by calculating the Wasserstein distances between two persistence diagrams, we can quantify the differences of persistence diagrams.

2.2 GSTH quantifies synchrony in coupled oscillators

Oscillations commonly occur in many organisms ranging from unicellular bacteria to mammalian cells. Temporal oscillations are thought to regulate metabolism, cell cycle, circadian rhythms, and rhythmic beating of the heart among many other physiological processes. We first validated GSTH by using it to quantify spatiotemporal patterns in the Kuramoto coupled oscillator model. This mathematical model is used to describe collective behavior in oscillating elements, such as the firing of neurons in the brain or activity of pacemaker cells in the heart [12–14]. The model consists of a group of oscillators, each with its own natural frequency, ω, that are coupled together and can influence each other's oscillations. The degree of synchronization among the oscillators is determined by a parameter called the coupling strength, R, which represents the strength of the interaction between the oscillators.

We simulated a network of 64 coupled oscillators with natural frequencies sampled from a normal distribution, $\omega_i \sim \mathcal{N}(1, \beta)$, and initial phase offsets sampled from a uniform distribution, $\theta_i \sim \mathcal{U}(0, 2\pi)$ for 200 timepoints with $\Delta t = 0.05$. We systematically varied the coupling strength, R, from 0 to 1 in our simulations. At low coupling strength, the oscillators act independently according to their natural frequency (Fig. 1A). The Kuramoto potential, a measure of synchrony in the system, remains close to 1.0 for the duration of the simulation, indicating the absence of synchronous behavior. The GSTH embeddings are temporally organized into continuous 1-D PHATE trajectories. The spacing between points in the trajectory is uniform and equal to the simulation time-step, as evident from the sharp decrease in the number of connected components at $\epsilon \approx \Delta t/2$ in the H_0 Betti curve.

At intermediate coupling strength the oscillators exhibit entrainment, characterized by decrease in phase variance and increase in phase-locking over time (Fig. 1B). At $R = 0.5$, the Kuramoto potential decreases from 1.0 and stabilizes at 0.3, indicating increase in synchrony over time. The spacing between points in the PHATE trajectory decreasing over time. Time points towards the end of the simulation are embedded close to each other, which indicates that the input signal, i.e. the phase of the oscillators, is converging over time. The rate of entrainment is measured by the H_0 Betti curve, where the number of connected components decreases rapidly at low values of the filtration parameter, $\epsilon \approx 0$, due to increase in nearest neighbor proximity between GSTH embeddings over time. At $R = 1.0$, the oscillators synchronize within 20 simulation time-steps, resulting in a 'shorter' PHATE trajectory where the majority of time points are embedded in close proximity (Fig. 1C). Compared to $R = 0.5$, the H_0 Betti curve for $R = 1.0$ decreases more rapidly, confirming a faster rate of entrainment at higher coupling strength. We are able to recovery of the coupling strength parameter from the PHATE embeddings with close to perfect accuracy. Although information is lost in the computation of Betti curves, coupling strength can still be recovered with a high accuracy using a supervised neural network (Table 1).

2.3 GSTH recovers parameters in intercellular Ca^{2+} signaling model

We used a 2D computational model of the *Drosophila* wing imaginal disc to demonstrate that GSTH can distinguish between intermittent and continuous patterns of Ca^{2+} signaling activity and recover model parameters from signaling data. This model accounts for the formation and degradation of IP_3, Ca^{2+} dynamics across the endoplasmic reticulum (ER), and IP_3R activation dynamics using a system of ordinary differential equations that model the dynamics of cytosolic IP_3, cytosolic Ca^{2+}, ER Ca^{2+} concentration, and the fraction of IP_3 receptors that have not been inactivated by Ca^{2+} [15].

To reproduce the spiking, transient, travelling wave and fluttering like activity observed in experiments, we systematically varied the V_{PLC} parameter, which governs the maximal production rate of IP_3, in individual cells of the simulated tissue. The type of signaling pattern was dependent on the number of cells in the tissue having a V_{PLC} value below, above, or equal to the Hopf bifurcation threshold for single-cells ($V_{\mathrm{PLC}} = 0.774$). Simulated cells that have a V_{PLC} value above the Hopf threshold, in the absence of agonist stimulation, are termed "initiator cells" and are poised to exhibit high levels of IP_3 production. Neighboring simulated cells with V_{PLC} values below the Hopf threshold are termed "standby cells" that receive a signal from initiator cells to propagate a signal. For instance, if a majority of standby cells have V_{PLC} values significantly below the critical Hopf bifurcation threshold, and standby cell V_{PLC} is randomly uniformly distributed between $0.1 - 0.5$, single-cell Ca^{2+} spikes occur only where initiator cells oscillate (Fig. 2A,i-iii). When we increased standby cell V_{PLC} values close to the lower end of the Hopf bifurcation point (Fig. 2B,i-iii), and standby cell V_{PLC} is randomly uniformly distributed between $0.25 - 0.60$, we noticed the formation of inter-cellular transients (ICTs). Finally, we observed the formation of inter-cellular calcium waves (ICWs) when standby cell V_{PLC} is randomly uniformly distributed between $0.4 - 0.8$, and fluttering phenotypes when standby cell V_{PLC} is randomly uniformly distributed between 1.4–1.5 (Fig. 2 C-D,i-iii), in cases where the majority of cells in the system were assigned a V_{PLC} close to or above the bifurcation threshold, thereby placing more cells in an initiator state. In the absence of initiator cells, Ca^{2+} activity was not observed.

GSTH analysis of the model simulations reveals two major trends. First, the GSTH embeddings generate 1-D PHATE trajectories that are organized by time. Gaps between successive time points in these trajectories indicate intermittent signaling behavior. A large number of gaps are present in the trajectories of spiking and ICT simulations, where a handful of initiator cells signal sporadically and the Ca^{2+} signal does not propagate via standby cells across the tissue (Fig. 2A-B,iv). By comparison, the trajectories of ICW and fluttering simulations are continuous and smooth, reflecting consistent and recurring signaling activity over time in the simulation (Fig. 2C-D,iv). Gaps in these trajectories are quantified by the persistent homology of connected components and visualized in the H_0 Betti curves. Gaps between points results in longer

persistence intervals. These intervals persist over large values of the filtration parameter, which is evident by the smaller slope of the H_0 Betti curves in simulations with intermittent signaling activity (Fig. 2A-D,iv). Another trend we observe is the presence of near circular, loop-like sections in the PHATE trajectory. These segments of the trajectory are indicative of quasi-periodicity in the input, i.e. returning close to a previously encountered state. For example, the fluttering simulations exhibit coordinated tissue-wide waves of Ca^{2+} activity that repeat over time (Fig. 2D). These repeating patterns of activity create persistent topological loops in the PHATE trajectory which are quantifed in the H_1 Betti curve. Note that similar, albeit discontinuous, loop-like structure is present in the spiking and ICT trajectories where there is no signaling activity in the tissue at multiple timepoints resulting in the creation of trivial loops.

2.4 GSTH identifies drug treatments using mouse epidermal Ca^{2+} signaling data

Having validated our method with computational models of cell signaling, we sought to understand signaling dynamics in an *in vivo* context. With the advent of genetically encoded calcium indicators (GECIs) [16] and innovations in live imaging techniques, investigators have begun to observe diverse patterns of Ca^{2+} signaling in many tissues types, however sophisticated analysis of Ca^{2+} signaling across broad spatial and temporal scales *in vivo* has been mainly limited to excitable tissues.

We explored the basal stem cell layer of the mouse epidermis as an *in vivo* model to study the spatiotemporal dynamics of Ca^{2+} signaling across a highly regenerative, non-excitable tissue. The mouse epidermis is a unique model system to study signaling dynamics, as it can be imaged without invasive surgery and signaling pathways can be manipulated via topical drug treatment. The dataset consisted of perturbations to various components of the Ca^{2+} signaling pathway to manipulate spatiotemporal Ca^{2+} dynamics, in comparison to homeostatic dynamics in the control condition [17]. We used GSTH to analyze signaling dynamics obtained from perturbations at two points of the Ca^{2+} signaling pathway via drug treatment. In order to block sarco/endoplasmic reticulum Ca^{2+}-ATPase (SERCA) activity and activate store operated Ca^{2+} entry (SOCE), mice were treated topically with thapsigargin dissolved in DMSO. We noticed an immediate increase in Ca^{2+} transients due to upregulation of SOCE. Conversely, we inhibited IP_3R activity by treating mice with 2-Aminoethoxydiphenyl borate (2-APB), leading to a global decrease in Ca^{2+} transients.

GSTH analysis based on image segmentation of the timelapse microscopy data revealed significant differences between the homeostatic dynamics in the control mice and the perturbed signaling dynamics in the drug-treated mice (Fig. 3). We consistently obtained smooth trajectories in the control condition indicating that Ca^{2+} activity steadily diffuses to neighboring cells in a directed and coordinated manner, similar to the intercellular calcium waves in simulations of the drosophila wing imaginal disc. TDA of these smooth trajectories revealed a few large scale loops appearing at high values of the spatial parameter, ϵ, suggesting quasi-periodic signaling activity over time. By comparison, we obtained noisy and discontinuous trajectories in the drug treatment conditions. Increase in transient activity due to thapsigargin treatment produces a tortuous trajectory with gaps and dense segments (Fig. 3A). The gaps indicate periods of low to zero activity followed by bursts of synchronous signaling. Decrease in transient activity due to 2-APB treatment generates a more disorganized and scattered signal embedding, similar to spiking behavior in the drosophila wing imaginal disc (Fig. 3B). A feed-forward neural network trained on PHATE trajectories as input features classified the control condition and drug treatments with an accuracy of 88%, beating the performance of the same network trained using features obtained from kymographs, CROCKERs and zig-zag persistent homology (Table 2).

2.5 GSTH reconstructs sinusoidal patterns shown to mice from visual cortex recordings

Using publicly available two-photon microscopy data [18], we showed that GSTH can recover sinusoidal grating patterns presented to mice from calcium imaging of layer 4 regions of the prefrontal visual cortex. Sinusoidal grating patterns are a type of visual stimulus consisting of parallel stripes that vary in brightness according to a sinusoidal function. These patterns are commonly used in vision research because they are simple, well-defined stimuli that can evoke robust responses in the visual cortex.

Moving striped patterns on a computer screen with varying spatial and temporal frequencies and orientations were shown to mice, fitted with head-plates and a cranial window. Each pattern was shown for 3 seconds with a 2-second break in between, and this was repeated 8 times for each pattern. The monitor was positioned in front of the mouse's eye and the images were collected from the layer 4 region in the visual cortex, at a rate of approximately 30 images per second.

We trained a 5-layer feed-forward neural network classifier to predict the ground-truth visual stimuli presented to the mice from GSTH trajectories of signaling in the visual cortex. We separately trained the classifier to predict variations in spatial frequency (SF), temporal frequency (TF) and orientation of the pattern (Table 2). In experiments involving variation in SF, we were able to predict the visual stimulus with 72% classification accuracy using 5-fold cross-validation. We performed better than a random classifier at predicting the orientation of the sinusoidal pattern, although the accuracy was significantly lower. Finally, we were unable to predict the temporal frequency of the sinusoidal pattern using GSTH trajectories, since our classifier was unable to perform better than random guessing. One possible explanation is that temporal frequency information is not represented in the layer 4 region of the visual cortex which was imaged in this study. Importantly, GSTH outperformed other methods of quantifying the signaling dynamics in all cases.

2.6 GSTH identifies oncogenic ERK signaling in epithelial breast cells *in vitro*

Having shown the utility of GSTH in analyzing large Ca^{2+} signaling datasets, we next investigated the dynamics of other signaling pathway proteins such as extracellular signal-regulated kinase (ERK), part of the Mitogen Activated Protein Kinase (MAPK) signaling network [19]. We took advantage of previously published data, where oncogenic epithelial cells induced ERK waves in wildtype neighboring cells [20]. We analyzed four datasets, in which cells from the chromosomally-normal human breast epithelial line, MCF10A, expressed the ERK Kinase Translocation Reporter (ERK KTR) [21] and a nuclear reporter H2B-iRFP, allowing for cell segmentation and measurement of ERK activity over time. Oncogenic doxycycline-inducible BRAF-V600E overexpression was achieved through lentiviral infection in cells (H2B-mClover nuclear marker) that were then cocultured with wild-type cells at a low fraction (1:100). When induced with Doxycycline, these mutant BRAF-V600E cells initiated ERK signaling in neighboring wildtype cells. Media-treated cocultures were included as a control. Aikin et al. have shown that ERK signaling in neighboring wildtype cells is dependent on expression of the membrane-tethered sheddase ADAM17 in oncogenic initiator cells [20]. Cocultures with doxycycline-induced and media-treated BRAF-V600E ADAM17 KO cells were therefore included as additional controls.

GSTH analysis of the ERK signaling data revealed similar trajectories for the No Dox, Dox-ADAM17 KO, ADAM17 KO control conditions (Fig. 4). No significant signaling activity was observed in these data, which resulted in scattered and discontinuous trajectories akin to spiking behaviour in the *Drosophila* wing imaginal disc model. The trajectory of Dox induced mutant BRAF-V600E cells, co-cultured with wildtype cells, was smooth in contrast.

2.7 Comparison to other methods

2.7.1 Time-delay embedding of cell signals and CROCKER

While kymographs can be useful summaries of spatiotemporal patterns in data, they are typically designed in a problem-specific manner and with some knowledge of the expected dynamics. A more agnostic and less application-specific is to quantify the shape of the underlying data manifold using topological data analysis. Specifically, we encode the persistent homology of cell signals by constructing a point cloud using Taken's time-delay embeddings, and transforming the point cloud into a Vietoris-Rips simplicial complex.

We encode the signaling activity at each cell using a 2D time-delay positional embedding, $(X(v_l, t_i - \delta), X(v_l, t_i))$. Given a population of cells with signaling activity changing over time, we then construct a Betti curve from the 2D point cloud of the time-delay embeddings of cells at each timepoint t_i, and concatenate these into a matrix. We discretize time, t, along the columns and the filtration parameter, ϵ, along the rows. Entries in the matrix are the kth dimensional Betti number, β_k, for specific values of (t, ϵ). The matrix is a topological signature of the time-varying data and it can be vectorized for input into machine learning algorithms. The matrix often visualized as a contour plot, lending to its name, the Contour Realization Of Computed k-dimensional hole Evolution in the Rips complex, or CROCKER [22].

2.7.2 Zig-zag persistent homology

Despite the utility of "ordinary" persistent homology in machine learning and data analysis contexts (see e.g. Hensel et al. [23] for a recent survey), the expressivity of such methods is limited by the requirement of data being nested along the filtration, i.e., along the filtering process. In some applications, the use of *zig-zag persistent homology* may thus be preferred. This notion of topological features is permits more general relations between the individual parts of the data. More precisely, given a data set, zig-zag persistent homology can calculate topological features along general sequences of nested data sets of the form

$$\mathrm{K}_1 \leftrightarrow \mathrm{K}_2 \cdots \mathrm{K}_{n-1} \leftrightarrow \mathrm{K}_n, \tag{7}$$

where "$\leftrightarrow$" may either denote subset or superset inclusions, i.e., inclusions of the form "$\hookleftarrow$" or "$\hookrightarrow$." In particular, the types of inclusions are allowed to vary; ordinary persistent homology with sublevel set inclusions is thus a special case of such a zig-zag filtration according to Equation 7. The name "zig-zag" denotes the common use case in which both types of inclusions alternate. This can be achieved by inserting the union of data sets, leading to

$$\mathrm{K}_1 \hookrightarrow \mathrm{K}_{1,2} \hookleftarrow \mathrm{K}_2 \hookrightarrow \cdots \hookrightarrow \mathrm{K}_{n-1,n} \hookleftarrow \mathrm{K}_n, \tag{8}$$

where $\mathrm{K}_{i-1,i} := \mathrm{K}_{i-1} \cup \mathrm{K}_i$. It is possible, in complete analogy to ordinary persistent homology, to track topological features along such a zig-zag filtration, making this formulation suitable particularly for time-varying data.

2.7.3 Kymographs

Kymographs are often used to graphically illustrate changes in the intensity of a signal over time in some spatial region. This plot is commonly used in the study of cell signaling, as it allows researchers to visualize and quantitatively analyze the dynamics of signaling pathways within cells.

The kymograph plot is typically constructed by dividing the tissue sample into a series of horizontal lines, or "stripes," that represent different time points. The intensity of the signal is then plotted along the vertical axis, with higher intensities representing stronger signaling activity.

By examining the kymograph plot, researchers can identify key characteristics of the signaling process, such as the rate at which the signal is transmitted, the strength of the signal at different time points, and the overall pattern of signaling activity within the tissue. This information can be used to better understand the role of signaling pathways in cell behavior and can provide insights into potential therapeutic targets for diseases related to aberrant signaling.

2.8 Ablation Study

2.8.1 PHATE trajectory ablation

We compare application of the proposed GSTH method on the three synthetic datasets with approaches that ablate or replace steps in the GSTH method. In particular, we test:

- Applying PHATE directly on the raw input signals to obtain time-trajectories, without the use of the geometric scattering transform.
- Applying PCA on the generated scattering coefficients instead of PHATE.
- Applying t-SNE on the generated scattering coefficients instead of PHATE.
- Applying UMAP on the generated scattering coefficients instead of PHATE.

For the synthetic testcase 1, we aim to compare the approaches for their stability to small perturbations as well as their ability to retrieve signal diffusion dynamics on the graph. As shown in Fig. 5A, after applying GSTH, time points with perturbed signals overlapped with time points with original signals, showing scattering transform and PHATE are invariant to small degrees of noise. The smooth trajectory also reflects that the scattering transform and PHATE of GSTH can effectively capture the signal propagation on the graph. By contrast, directly using PHATE on the raw input signals will result in the condensed timepoints in Fig. 5C, thus failing to retrieve the dynamics. While applying PCA (Fig. 5D) and t-SNE (Fig. 5E) on the generated scattering coefficients can retrieve the dynamics to some extent, Fig. 5D shows a more disrupted trajectory and the trajectory from Fig. 5E overlaps with itself. Similarly, applying UMAP (Fig. 5F) on the generated scattering coefficients also led to overlapping timepoints. All these methods thus failed to reflect the propagation of signals on the graph.

For the second synthetic dataset, we further compare the ability of different approaches to retrieve signal diffusion dynamics on the graph under a more complex scenario. For GSTH (Fig. 5H) time points from two signal sources formed two branches with their starting points near each other in PHATE coordinates. Thus from one end to the next this is akin to a signal condensing and then diffusing again. As expected, this creates a loop-like structure in the PHATE graph. However, directly applying PHATE on the raw signals (Fig. 5J) results in multiple scattered points separated from the main trajectory, demonstrating that using PHATE only is not able to fully capture and distinguish the signals. Furthermore, although applying PCA on the scattering coefficients (Fig. 5K) generates two separate trajectories, they fail to form the loop-like structure as with using GSTH. Applying t-SNE (Fig. 5L) and UMAP (Fig. 5M) on the generated scattering coefficients also failed to form loop-like structures.

Finally, for the third synthetic dataset, we aim to simulate the propagation of signals similar to that observed in epithelial cells. This will help us to better understand what types of signals GSTH can capture. The propagation of signals among cells can reflect the communication within groups of cells (corresponding to the clusters of Ca^{2+} transients among epithelial cells using Ca^{2+}-sensor imaging), while each cell itself can also generate independent signals without communicating with other cells (corresponding to single cells spiking using Ca^{2+} fluorescence imaging). As in Figure 5O for GSTH, the time points formed smooth trajectories at first simulating the propagation of signals on the graph, then disrupted trajectories corresponding to random spiking of single cells. In comparison, using PHATE directly on the raw input signals (Fig. 5Q) results in more dense clustering (for the initial stage when the signal is just diffusing on the graph) and using t-SNE on the scattering coefficients generates more scattered clustering, making it hard to identify the inner dynamics (Fig. 5S). Although applying PCA (Fig. 5R) and UMAP (Fig. 5T) on the scattering coefficients can reflect part of the dynamics, they also generate very condensed trajectories for the early stage, when the signal is simply diffusing on the graph.

In addition, we computed the Wasserstein distances between the persistence diagram from our GSTH method and persistence diagrams from other methods using the three synthetic datasets (Fig. 5G, 5N, 5U). We showed that the persistence diagrams from our GSTH method are different from persistence diagrams produced using other methods ablating different parts of GSTH and visualized the distances with heatmaps (Fig. 5G, 5N, 5U).

3 Methods

3.0.1 Cellular Graphs and Graph Signals

We represent the imaged tissue as a graph $G = \{V, E\}$, consisting of nodes $v_i \in V$ and edges $(v_j, v_k) \in E$, where each node v_i represents a cell and a pair of nodes v_j and v_k is connected with an edge based on a predefined criterion. For epithelial cells, we connect nodes that are spatially adjacent (within 2 μm of each other), as the flow of signals is thought to be between spatially proximal cells. On the other hand, neurons can have long processes that are often hard to image, and therefore we use correlation between neurons' Ca^{2+} signals to connect the neuronal graph. Finally, the connectivity of graph G can be described by its adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij} = 1$ if v_i and v_j are connected and 0 otherwise. The degree of each vertex is defined as a diagonal matrix $\mathbf{D}$, where $\mathbf{D}_{ii} = \sum_{j=1}^{i} \mathbf{A}_{ij}$.

Graph signals can associate with each node or edge in a graph. In the Ca^{2+} signaling data, the signals associated with cell v_i is the normalized Ca^{2+} fluorescence intensity at each timestep t. Since every cell has a related Ca^{2+} signal, this signal $X(v_i, t)$ is defined over the whole graph for timestep t.

3.0.2 Diffusion Geometry

A useful assumption in representation learning is that high dimensional data originates from an intrinsic low dimensional manifold that is mapped via nonlinear functions to observable high dimensional measurements; this is commonly referred to as the manifold assumption. Formally, let $\mathcal{M}^d$ be a hidden d-dimensional manifold that is only observable via a collection of $n \gg d$ nonlinear functions $f_1, \ldots, f_n : \mathcal{M}^d \to \mathbb{R}$ that enable its immersion in a high dimensional ambient space as $F(\mathcal{M}^d) = \{\mathbf{f}(z) = (f_1(z), \ldots, f_n(z))^T : z \in \mathcal{M}^d\} \subseteq \mathbb{R}^n$ from which data is collected. Conversely, given a dataset $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ of high dimensional observations, manifold learning methods assume data points originate from a sampling $Z = \{z_i\}_{i=1}^N \in$

$\mathcal{M}^d$ of the underlying manifold via $x_i = \mathbf{f}(z_i)$, $i = 1, \ldots, n$, and aim to learn a low dimensional intrinsic representation that approximates the manifold geometry of $\mathcal{M}^d$.

To learn a manifold geometry from collected data, scientists often use the diffusion maps construction of [24] that uses diffusion coordinates to provide a natural global coordinate system derived from eigenfunctions of the heat kernel, or equivalently the Laplace-Beltrami operator, over manifold geometries. This construction starts by considering local similarities defined via a kernel $\mathcal{K}(x, y)$, $x, y \in F(\mathcal{M}^d)$, that captures local neighborhoods in the data. We note that a popular choice for $\mathcal{K}$ is the Gaussian kernel $\exp(-\|x-y\|^2/\sigma)$, where $\sigma > 0$ is interpreted as a user-configurable neighborhood size. However, such neighborhoods encode sampling density information together with local geometric information. To construct a diffusion geometry that is robust to sampling density variations, we use an anisotropic kernel:

$$\mathcal{K}(x, y) = \frac{\mathcal{G}(x, y)}{\|\mathcal{G}(x, \cdot)\|_1^{\alpha} \|\mathcal{G}(y, \cdot)\|_1^{\alpha}}, \text{ where } \quad \mathcal{G}(x, y) = e^{-\frac{\|x-y\|^2}{\sigma}}, \tag{9}$$

as proposed in [24], where $0 \leq \alpha \leq 1$ controls the separation of geometry from density, with $\alpha = 0$ yielding the classic Gaussian kernel, and $\alpha = 1$ completely removing density and providing a geometric equivalent to uniform sampling of the underlying manifold. Next, the similarities encoded by $\mathcal{K}$ are normalized to define transition probabilities $p(x, y) = \frac{\mathcal{K}(x,y)}{\|\mathcal{K}(x,\cdot)\|_1}$ that are organized in an $N \times N$ row stochastic matrix:

$$\mathbf{P}_{ij} = p(x_i, x_j) \tag{10}$$

that describes a Markovian diffusion process over the intrinsic geometry of the data. Finally, a diffusion map [24] is defined by taking the eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ and (corresponding) eigenvectors $\{\phi_j\}_{j=1}^N$ of $\mathbf{P}$, and mapping each data point $x_i \in X$ to an N dimensional vector $\Phi_t(x_i) = [\lambda_1^t \phi_1(x_i), \ldots, \lambda_N^t \phi_N(x_i)]^T$, where t represents a diffusion-time (i.e., number of transitions considered in the diffusion process). In general, as t increases, most of the eigenvalues λ_j^t, $j = 1, \ldots, N$, become negligible, and thus truncated diffusion map coordinates can be used for dimensionality reduction [24].

3.0.3 PHATE

PHATE is a dimensionality reduction method that captures both local and global nonlinear structure through constructing a diffusion geometry [9]. It computes the diffusion operator as in Equation 10. However, rather than eigendecomposing this operator to find new coordinates, PHATE creates a new distance matrix from $\mathbf{P}$ by defining an M-divergence between datapoints, called *potential distance* as $\mathbf{ID}_{i,j} = \| \log \mathbf{P}_{i,:}^t - \log \mathbf{P}_{j,:}^t \|_2$ between corresponding t-step diffusion probability distributions of the two points.

The advantage of this step is that the information theoretic distance between probabilities emphasizes differences in lower probabilities (corresponding to distant points) as well as high probabilities (corresponding to neighbors), and therefore globally contextualizes the point. The resulting information distance matrix **ID** is finally embedded into a low dimensional (2D or 3D) space by metric multidimensional scaling (MDS), and makes it possible to visualize intrinsic geometric information from data. In [9], authors demonstrate that PHATE performs better than all compared methods including diffusion maps and UMAP in preserving *denoised manifold affinity (DeMAP)* in low dimensions and, in particular, excels at preserving trajectory structures without shattering.

3.0.4 Persistent homology and topological data analysis

Topological data analysis (TDA) refers to techniques for understanding complex datasets by their topological features, i.e., their connectivity [25]. Here we focus on the topological features of a data graph where the simplest set of topological features are given by the number of connected components b_0 and the number of cycles b_1, respectively. Such counts, also known as the Betti numbers, are coarse graph descriptors that are invariant under graph isomorphisms. Their expressivity is increased by considering a function $f\colon V \times V \to \mathbb{R}$ on the vertices of a graph, $G = (V, E)$, with vertex set V and edge set E. Since V has finite cardinality, so does the image $\mathrm{im} f$, i.e., $\mathrm{im} f = \{w_1, w_2, \ldots, w_n\}$.

Without loss of generality, we assume that $w_1 \leq \cdots \leq w_n$. We write G_i for the subgraph induced by filtering according to w_i, such that the edges satisfy $E_i := \{(u, v) \in E \mid \|u - v\|_2^2 \leq w_i\}$. The subgraphs G_i satisfy a nesting property, as $G_1 \subseteq G_2 \subseteq \cdots \subseteq G_n$. When analyzing a point cloud, the vertices of each G_i arise from spatial coordinates for the data and w_i constitutes a distance threshold between points, such that G_n is a fully-connected graph, containing all the vertices from V. This is commonly known as the Vietoris-Rips (VR) filtration.

It is then possible to calculate topological features alongside this *filtration* of graphs, tracking their appearance and disappearance as the graph grows. If a topological feature is created in G_i, but destroyed in G_j (it might be destroyed because two connected components merge, for instance), we represent this by storing the point (w_i, w_j) in the *persistence diagram* $\mathcal{D}_f$ associated to G. Another simple descriptor is given by the Betti curve of dimension d of a diagram $\mathcal{D}$, which refers to the sequence of Betti numbers of dimension d in $\mathcal{D}$, evaluated for each threshold w_i.

3.1 Mathematical models and synthetic data

3.1.1 Kuramoto model of weakly coupled oscillators

The Kuramoto model [26] describes a weakly coupled system of oscillators, with dynamics governed by:

$$\dot{\phi_k} = \omega_k + \frac{\epsilon}{N} \sum_{j=1}^{N} \sin(\phi_j - \phi_k + \alpha_k) \tag{11}$$

where ω_k are the natural frequencies and α_k are the phase shift parameters.

By using the complex mean-field, $Z = Re^{i\theta} = {}^1\!/_N \sum_j e^{i\phi_j}$, the system can be reduced to:

$$\dot{\phi_k} = \omega_k + \epsilon R \sin(\theta - \phi_k + \alpha_k) \tag{12}$$

The order parameter R describes the degree of synchrony, $R = 1$ denotes full synchronization and at $R = 0$ each oscillator acts independently.

3.1.2 Intercellular Ca^{2+} signaling model

Ca^{2+} signaling has been implicated as a key molecular pathway in developing and wounded epithelial tissues [27–30]. The *Drosophila* wing imaginal disc is a classic system in which to explore intercellular communi-

cation in a developmental context where cells must highly coordinate their behaviors [27, 28, 31, 32]. We ran simulations of a generalized multicellular model of Ca^{2+}, developed by [15], that captures all the key categories of Ca^{2+} activity in nonexcitable tissues. The model simulates two different populations of cells, initiator and standby cells, within tissues that are connected by gap junction proteins. The initiator cells have elevated levels of Phospholipase C (PLC) activity, and standby cells exhibit baseline activity. The model reproduced single-cell Ca^{2+} spikes upon insulin stimulation and intercellular Ca^{2+} waves dependent on Gαq activity.

3.2 Experimental datasets

3.2.1 Epidermal Ca^{2+} signaling dataset

In vivo imaging was performed on mice with a Ca^{2+} sensor expressed in all epidermal cells (*K14-Cre; Rosa26-CAG-LSL-GCaMP6s*). In order to block SERCA (sarco/endoplasmic reticulum Ca^{2+}-ATPase) activity and activate store operated Ca^{2+} entry (SOCE), mice were treated with thapsigargin or DMSO vehicle control (see Biological Methods for details). Conversely, mice were treated with 2-Aminoethoxydiphenyl borate (2-APB) to inhibit IP_3R. In order to segment all cells in the field of view, including non-flashing cells, we used part of the MATLAB (version R2018b) package from Romano et al. [33], a watershed segmentation method. We normalized the fluorescence intensity of each cell at each timepoint to the minimum fluorescence intensity of that cell as a baseline.

3.2.2 Prefrontal visual cortex dataset

Image data generated in [34] were downloaded from the Collaborative Research in Computational Neuroscience (CRCNS) data sharing website. The image data were acquired using a two-photon microscope and images were collected at approximately 30 Hz with 512 lines per frame. Adult mice were fitted with head plates and a cranial window and allowed to run on a freely rotating disc while visual stimuli were presented to the eye contralateral to the imaged region (layer 4 cells of the primary visual cortex). Drifting sinusoidal gratings were shown in random order for all combinations of 5 spatial frequencies, 5 temporal frequencies, and 8 orientations (0 to 315 degrees in 45 degree steps). These stimuli were presented 8 times each for 3 seconds.

3.2.3 ERK signaling dataset

Raw data from [20] were shared by the authors. In all conditions, a cell reporter line derived from the chromosomally-normal human breast epithelial line, MCF10A, expressed the ERK Kinase Translocation Reporter (ERK KTR) [21] and a nuclear reporter H2B-iRFP, allowing for cell segmentation and measurement of ERK activity over time. Oncogenic doxycycline-inducible BRAF-V600E overexpression was achieved through lentiviral infection in cells (H2B-mClover nuclear marker) that were then cocultured with wild-type cells at a low fraction (1:100). When induced with Doxycycline, these mutant BRAF-V600E cells initiated ERK signaling in neighboring wildtype cells. Media-treated cocultures were included as a control. Aikin et al. have shown that ERK signaling in neighboring wildtype cells is dependent on expression of the membrane-tethered sheddase ADAM17 in oncogenic initiator cells. Cocultures with doxycycline-induced and media-treated BRAF-V600E ADAM17 KO cells were therefore included as additional controls. This constituted four different experimental conditions that we were able to compare using GSTH. Images were

acquired every 5 minutes for 181 timepoints. 10 positions were imaged per condition and > 1500 cells were quantified per position.

3.3 Biological Methods

3.3.1 Experimental setup for epidermal dataset

K14-Cre [35] mice were obtained from E. Fuchs (Rockefeller University). K14-H2BmCherry mice were generated in the laboratory and described previously[36]. Rosa26-CAG-LSL-GCaMP6s [37] mice were obtained from The Jackson Laboratory. Mice with a Ca^{2+} sensor expressed in all epidermal cells (*K14-Cre; Rosa26-CAG-LSL-GCaMP6s*) were treated with thapsigargin, 2-Aminoethoxydiphenyl borate (2-APB), or DMSO vehicle control. Thapsigargin was dissolved in a 25 mg ml^{-1} stock solution in dimethyl sulfoxide (DMSO) and then diluted 100 times in 100% petroleum jelly (Vaseline; final concentration is 250 mg ml $^{-1}$). One hundred micrograms of the mixture of thapsigargin and petroleum jelly was spread evenly on the ear of the anesthetized mouse 1 hour before imaging. 2-APB was dissolved in a 25 mg ml^{-1} stock solution in dimethyl sulfoxide (DMSO) and then diluted 100 times in 100% petroleum jelly (Vaseline; final concentration is 250 mg ml $^{-1}$). A mixture of 100% DMSO in petroleum jelly (1:100) was used as a vehicle control. Mice from experimental and control groups were randomly selected from either sex for live imaging experiments. No blinding was done. All procedures involving animal subjects were performed under the approval of the Institutional Animal Care and Use Committee (IACUC) of the Yale School of Medicine.

Imaging procedures were adapted from those previously described [38, 39]. All imaging was performed in distal regions of the ear skin during prolonged telogen, with hair removed using depilatory cream (Nair) at least 2 days before the start of each experiment. Mice were anaesthetized using an isoflurane chamber and then transferred to the imaging stage and maintained on anesthesia throughout the course of the experiment with vaporized isoflurane delivered by a nose cone (1.25% in oxygen and air). Mice were placed on a warming pad during imaging. The ear was mounted on a custom-made stage and a glass coverslip was placed directly against it. Image stacks were acquired with a LaVision TriM Scope II (LaVision Biotec) laser scanning microscope equipped with a tunable Two-photon Vision II Ti:Sapphire (Coherent) Ti:Sapphire laser and tunable Two-photon Chameleon Discovery Ti:Sapphire laser (Coherent) and Imspector Pro (LaVision Biotec, v.7.0.129.0). To acquire serial optical sections, a laser beam (940 nm, 1120 nm for mice and whole-mount staining) was focused through a 20x or 40x water-immersion lens (NA 1.0 and 1.1 respectively; Zeiss) and scanned with a field of view of 500 μm^2 or 304 μm^2, respectively at 800 Hz or through a 25x water-immersion lens (NA 1.0; Nikon) and scanned with a field of view of 486 μm^2 at 800 Hz. Visualization of collagen was achieved via the second harmonic signal at 940 nm. For all time-lapse movies, the live mouse remained anesthetized for the length of the experiment and serial optical sections were captured at intervals of 2 seconds.

3.3.2 Experimental setup for prefrontal visual cortex dataset

On postnatal day 36, the mouse was injected with AAV1.Syn.Flex.GCaMP6s.WPRE lot CS0215 produced by the University of Pennsylvania Vector Core and designed by the GENIE Project at Janelia Farms. At postnatal day 59, the mouse underwent surgery for head-plate implantation and craniotomy for glass window placement over the visual cortex. During the experiment, the mouse was habituated to head-fixation and the imaging rig for 3 weeks before visual stimuli were presented. The visual stimuli consisted of drifting sinusoidal gratings presented at different frequencies and orientations on an LCD monitor. The mouse's eye was positioned 22 cm away from the center of the monitor. Image data were collected using a custom-built

two-photon microscope, and images were acquired at approximately 30 Hz with 512 lines per frame. The image area corresponds to layer 4 in the windowed adult mice, and the experiment was designed to track eye and motion movements while visual stimuli were presented.

4 Conclusion

In conclusion, this study demonstrates the remarkable potential of combining geometric scattering and persistent homology for the quantification and analysis of spatiotemporal signaling in various biological systems. Through the successful prediction of synchrony in coupled oscillators and the recovery of model parameters from simulations of calcium signaling in the Drosophila wing imaginal disc, the proposed framework has showcased its capability to capture intricate patterns and reveal hidden information within complex cellular processes.

Furthermore, GSTH's ability to classify drug treatments by analyzing calcium signals in the mouse epidermis and identify oncogenic ERK signaling in epithelial breast cells highlights the framework's potential for practical applications, such as drug discovery and personalized medicine. Additionally, the recovery of sinusoidal visual stimuli presented to mice from calcium imaging of the visual cortex signifies the framework's capacity to bridge the gap between cellular-level events and organism-level responses, a crucial aspect for understanding the intricate interplay between cellular signaling and organism behavior.

Overall, the integration of geometric scattering and persistent homology within a single framework offers a powerful and versatile tool for studying spatiotemporal patterns in cell signaling. The demonstrated success across various biological systems and applications underscores the potential of this framework in advancing our understanding of cellular communication, coordination, and regulation. By unlocking new insights into the complex dynamics of cellular communication, this innovative approach holds great promise for advancing our understanding of diverse biological systems and has the potential to transform the fields of systems biology, drug development, and neuroscience, among others.

Data and code availability

The source code is available at `https://github.com/krishnaswamylab/GSTH`.

References

1. Basson, M. A. Signaling in Cell Differentiation and Morphogenesis. *Cold Spring Harbor Perspectives in Biology* **4,** a008151. ISSN: 1943-0264. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367549/` (2023) (June 2012).
2. Perrimon, N., Pitsouli, C. & Shilo, B.-Z. Signaling Mechanisms Controlling Cell Fate and Embryonic Patterning. en. *Cold Spring Harbor Perspectives in Biology* **4.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, a005975. ISSN: , 1943-0264. `http://cshperspectives.cshlp.org/content/4/8/a005975` (2023) (Aug. 2012).

3. Eming, S. A., Martin, P. & Tomic-Canic, M. Wound repair and regeneration: Mechanisms, signaling, and translation. *Science translational medicine* **6,** 265sr6. ISSN: 1946-6234. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4973620/ (2023) (Dec. 2014).

4. Ghilardi, S. J., O'Reilly, B. M. & Sgro, A. E. Intracellular signaling dynamics and their role in coordinating tissue repair. eng. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* **12,** e1479. ISSN: 1939-005X (May 2020).

5. Hotamisligil, G. S. & Davis, R. J. Cell Signaling and Stress Responses. en. *Cold Spring Harbor Perspectives in Biology* **8.** Publisher: Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5046695/ (2023) (Oct. 2016).

6. Sung, M.-H. & Hager, G. L. Nonlinear Dependencies of Biochemical Reactions for Context-specific Signaling Dynamics. en. *Scientific Reports* **2.** Number: 1 Publisher: Nature Publishing Group, 616. ISSN: 2045-2322. https://www.nature.com/articles/srep00616 (2023) (Aug. 2012).

7. Gao, F., Wolf, G. & Hirn, M. *Geometric Scattering for Graph Data Analysis* in *Proceedings of the 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, Sept. 2019), 2122–2131. http://proceedings.mlr.press/v97/gao19e.html.

8. Gama, F., Ribeiro, A. & Bruna, J. *Diffusion Scattering Transforms on Graphs* 2018. https://arxiv.org/abs/1806.08829.

9. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* **37,** 1482–1492 (Dec. 2019).

10. Perlmutter, M., Gao, F., Wolf, G. & Hirn, M. *Geometric Wavelet Scattering Networks on Compact Riemannian Manifolds* in *Proceedings of The First Mathematical and Scientific Machine Learning Conference* (eds Lu, J. & Ward, R.) **107** (PMLR, 20–24 Jul 2020), 570–604. https://proceedings.mlr.press/v107/perlmutter20a.html.

11. Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. Stability of persistence diagrams. *Discrete & computational geometry* **37,** 103–120 (2007).

12. Kuramoto, Y. International symposium on mathematical problems in theoretical physics. *Lecture notes in Physics* **30,** 420 (1975).

13. Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Spigler, R. The Kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of Modern Physics* **77.** Publisher: American Physical Society, 137–185. https://link.aps.org/doi/10.1103/RevModPhys.77.137 (2022) (Apr. 2005).

14. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. en. *Physica D: Nonlinear Phenomena* **143,** 1–20. ISSN: 0167-2789. https://www.sciencedirect.com/science/article/pii/S0167278900000944 (2022) (Sept. 2000).

15. Soundarrajan, D. K., Huizar, F. J., Paravitorghabeh, R., Robinett, T. & Zartman, J. J. From spikes to intercellular waves: Tuning intercellular calcium signaling dynamics modulates organ size control. en. *PLOS Computational Biology* **17.** Publisher: Public Library of Science, e1009543. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009543 (2022) (Nov. 2021).

16. Miyawaki, A. *et al.* Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature* **388,** 882–887. ISSN: 00280836. https://www.nature.com/articles/42264 (1997).

17. Moore, J. L. *et al. Tissue-wide coordination of calcium signaling regulates the epithelial stem cell pool during homeostasis* en. Pages: 2021.10.12.464066 Section: New Results. Oct. 2021. https://www.biorxiv.org/content/10.1101/2021.10.12.464066v2 (2023).

18. Garner, A. *In vivo calcium imaging of layer 4 cells in the mouse using sinusoidal grating stimuli* 2014.

19. Lavoie, H., Gagnon, J. & Therrien, M. ERK signalling: a master regulator of cell behaviour, life and fate. *Nature reviews. Molecular cell biology* **21,** 607–632. ISSN: 1471-0072. https://doi.org/10.1038/s41580-020-0255-7 (Oct. 2020).

20. Aikin, T. J., Peterson, A. F., Pokrass, M. J., Clark, H. R. & Regot, S. MAPK activity dynamics regulate non-cell autonomous effects of oncogene expression. *eLife* **9** (eds Rosenblatt, J., Cooper, J. A. & Rosenblatt, J.) Publisher: eLife Sciences Publications, Ltd, e60541. ISSN: 2050-084X. https://doi.org/10.7554/eLife.60541 (2022) (Sept. 2020).

21. Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S. & Covert, M. W. High-Sensitivity Measurements of Multiple Kinase Activities in Live Single Cells. *Cell* **157,** 1724–1734. ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S0092867414006552 (2014).

22. Topaz, C. M., Ziegelmeier, L. & Halverson, T. Topological Data Analysis of Biological Aggregation Models. en. *PLOS ONE* **10.** Publisher: Public Library of Science, e0126383. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0126383 (2022) (May 2015).

23. Hensel, F., Moor, M. & Rieck, B. A Survey of Topological Machine Learning Methods. *Frontiers in Artificial Intelligence* **4.** ISSN: 2624-8212 (2021).

24. Coifman, R. R. & Maggioni, M. Diffusion wavelets. *Applied and Computational Harmonic Analysis* **21,** 53–94 (2006).

25. Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **46,** 255–308 (2009).

26. Pikovsky, A. & Rosenblum, M. Dynamics of globally coupled oscillators: Progress and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **25.** Publisher: American Institute of Physics, 097616. ISSN: 1054-1500. https://aip.scitation.org/doi/10.1063/1.4922971 (2022) (Sept. 2015).

27. Restrepo, S. & Basler, K. Drosophila wing imaginal discs respond to mechanical injury via slow InsP3R-mediated intercellular calcium waves. *Nature Communications* **7,** 1–9. ISSN: 20411723. www.nature.com/naturecommunications (Aug. 2016).

28. Balaji, R. *et al.* Calcium spikes, waves and oscillations in a large, patterned epithelial tissue. *Scientific Reports* **7,** 42786. ISSN: 2045-2322. http://www.nature.com/articles/srep42786 (Feb. 2017).

29. Ohno, Y. & Otaki, J. M. Spontaneous long-range calcium waves in developing butterfly wings. *BMC Developmental Biology* **15,** 17. ISSN: 1471213X. http://www.biomedcentral.com/1471-213X/15/17 (Mar. 2015).

30. Peters, M. A., Teramoto, T., White, J. Q., Iwasaki, K. & Jorgensen, E. M. A Calcium Wave Mediated by Gap Junctions Coordinates a Rhythmic Behavior in C. elegans. *Current Biology* **17,** 1601–1608. ISSN: 09609822 (2007).

31. Buchmann, A., Alber, M. & Zartman, J. J. Sizing it up: The mechanical feedback hypothesis of organ growth regulation. *Seminars in Cell & Developmental Biology* **35.** Regulated Necrosis & Modeling developmental signaling pathways & Development of connective maps in the brain, 73–81. ISSN: 1084-9521. https://www.sciencedirect.com/science/article/pii/S1084952114001918 (2014).

32. Restrepo, S., Zartman, J. J. & Basler, K. Coordination of Patterning and Growth by the Morphogen DPP. *Current Biology* **24,** R245–R255. ISSN: 0960-9822. https://www.sciencedirect.com/science/article/pii/S0960982214001201 (2014).

33. Romano, S. A. *et al.* An integrated calcium imaging processing toolbox for the analysis of neuronal population dynamics. *PLoS Computational Biology* **13.** ISSN: 15537358. https://pubmed.ncbi.nlm.nih.gov/28591182/ (June 2017).

34. Garner, A. In vivo calcium imaging of layer 4 cells in the mouse using sinusoidal grating stimuli. (2014).

35. Vasioukhin, V., Degenstein, L., Wise, B. & Fuchs, E. The magical touch: Genome targeting in epidermal stem cells induced by tamoxifen application to mouse skin. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 8551–8556. ISSN: 00278424. www.pnas.org. (July 1999).

36. Mesa, K. R. *et al.* Niche-induced cell death and epithelial phagocytosis regulate hair follicle stem cell pool. *Nature* **522,** 94–97. ISSN: 1476-4687. http://www.nature.com/nature/journal/v522/n7554/fp/nature14306_ja.html (2015).

37. Madisen, L. *et al.* Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron* **85,** 942–958. ISSN: 10974199 (Mar. 2015).

38. Pineda, C. M. *et al.* Intravital imaging of hair follicle regeneration in the mouse. *Nature Protocols* **10,** 1116–1130. ISSN: 1754-2189. http://www.nature.com/doifinder/10.1038/nprot.2015.070 (June 2015).

39. Rompolas, P. *et al.* Live imaging of stem cell and progeny behaviour in physiological hair-follicle regeneration. *Nature* **487,** 496–499. ISSN: 0028-0836 (2012).

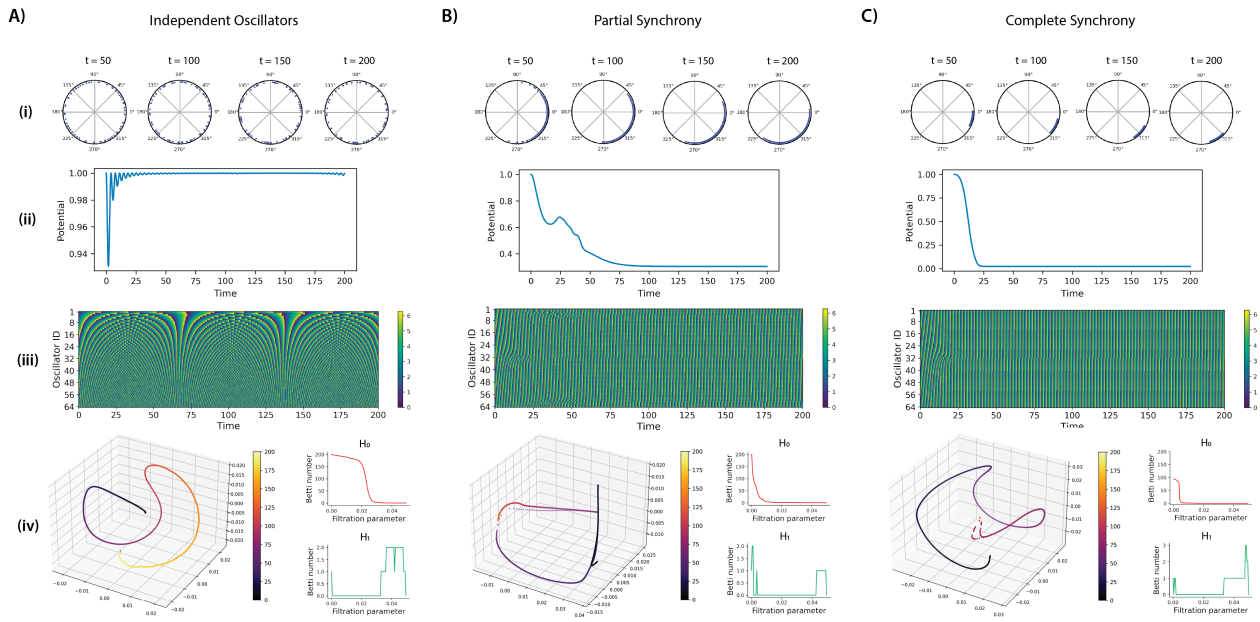

Figure 1: **GSTH captures synchrony and phase locking in the Kuramoto model at varying coupling strengths.** Phase plot (i), potential function (ii), kymograph (iii), PHATE trajectory and betti curves (iv) of the Kuramoto model at different coupling strengths. (**(A)**), Independent oscillations at low coupling strength ($R = 0.1$) **(B)** PHATE trajectories and betti curves at $R = 0.5$ (partial synchrony). **(C)** PHATE trajectories and betti curves at $R = 0.01$ (no synchrony).

Table 1: Comparison of GSTH with other methods based on MSE for model parameter recovery, lower is better. 10 simulations were performed for each parameter value. 5 parameter values were sampled linearly from the range provided.

Model	Parameter	Range	MSE			
			GSTH	Kymograph	CROCKER	Zigzag
Kuramoto	R	$0.1 - 1.0$	**0.0081**	N/A	0.0174	0.0193
ICW	K_{SERCA}	$0.09 - 0.15\ \mu M$	**0.0146**	0.0384	0.0476	0.0515
	V_{SERCA}	$0.675 - 0.99\ \mu M s^{-1}$	**0.0646**	0.0187	0.0639	0.0582
	$\tau_{\max}$	$400 - 1200\ s^{-1}$	24.81	**19.72**	30.12	32.71

Table 2: Comparison of GSTH with other methods based on classification accuracy on experimental datasets, higher is better.

Exp. Dataset	Num. Classes	Range	Classification Accuracy			
			GSTH	Kymograph	CROCKER	Zigzag PH
Epidermal	3	DMSO (control), thapsigargin, 2-APB	**0.88**	0.76	0.71	0.64
ERK	4	WT, $\text{ADAM17}^{\text{KO}}$ (w/wo Dox treatment)	**0.65**	0.49	0.42	0.38
	5 (SF)	$0.02 - 0.32$ Hz	**0.72**	N/A	0.68	0.59
PVC-7	5 (TF)	$1, 2, 4, 8, 15$ Hz	**0.23**	N/A	0.19	0.17
	8 (orient.)	$0 - 315$ deg	**0.45**	N/A	0.36	0.28

Table 3: Performance of GSTH model ablations on model parameter recovery (MSE) and classification of experimental conditions.

		Mathematical Model[1]				Experimental Data[2]		
		Kuramoto	Intercellular Ca^{2+} Waves			Epidermal	ERK	PVC
Model	Ablation	R	K_{SERCA}	V_{SERCA}	$\tau_{\max}$	Signaling	Signaling	Signaling
GSTH	None	0.0063	0.0372	0.0646	24.81	0.89	0.65	0.44
GST	PH	**0.0016**	**0.0189**	**0.0232**	**9.74**	**0.95**	0.73	**0.62**
GST-RNN	PH	0.0058	0.0417	0.0722	26.94	0.86	0.58	0.37
GST-κ	PH	0.0347	N/A	0.1953	N/A	N/A	N/A	N/A
GS-PCA-H	PHATE	0.0125	0.0506	0.1103	51.63	0.77	0.57	0.44
GS-tSNE-H	PHATE	0.0118	0.0438	0.1078	38.29	0.81	0.73	0.51
Graph PH	GST	0.0124	0.0397	0.0859	11.32	0.91	**0.76**	

[a]Mean squared error (MSE) for regression, lower is better.
[b]Classification accuracy, higher is better.

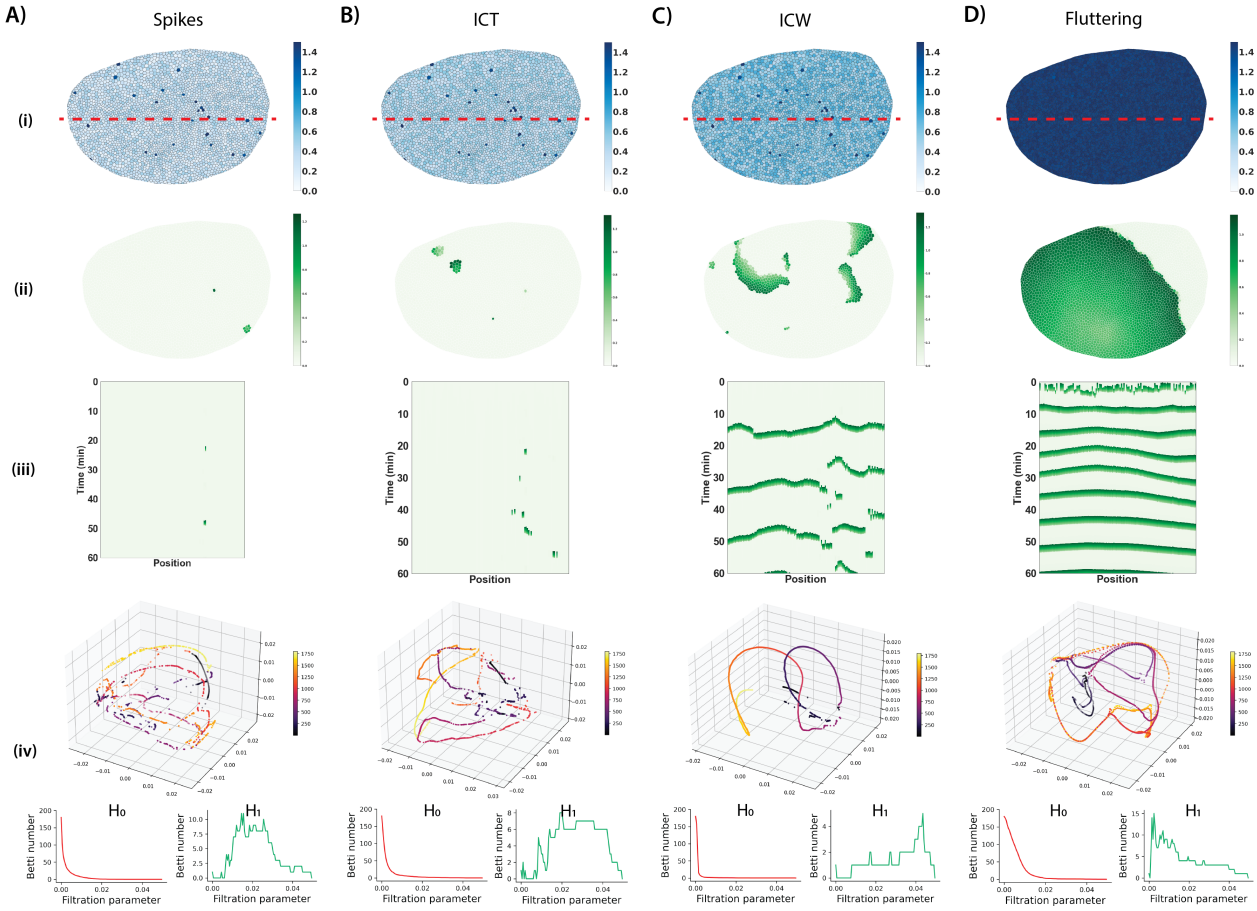

Figure 2: **GSTH captures four distinct patterns of Ca^{2+} signaling activity in developing Drosophila wing imaginal discs.** Spatiotemporal patterns in simulations recapitulating the key classes of multicellular Ca^{2+} activity are analyzed using GSTH. A dashed line through the A/P direction (red) demonstrates where the kymograph line is drawn that produces the simulated tissue's corresponding kymograph. When the majority of cells have V_{PLC} values below the Hopf bifurcation threshold, single-cell Ca^{2+} spikes are seen **(A)**. Intercellular Ca^{2+} transients (ICT) are observed **(B)** as the distribution of VPLC in cells is increased. Periodic intercellular Ca^{2+} waves (ICW) emerge as V_{PLC} is increased further **(C)**. "Fluttering" occurs when V_{PLC} levels in all of the cells in the disc is above Hopf bifurcation threshold **(D)**.

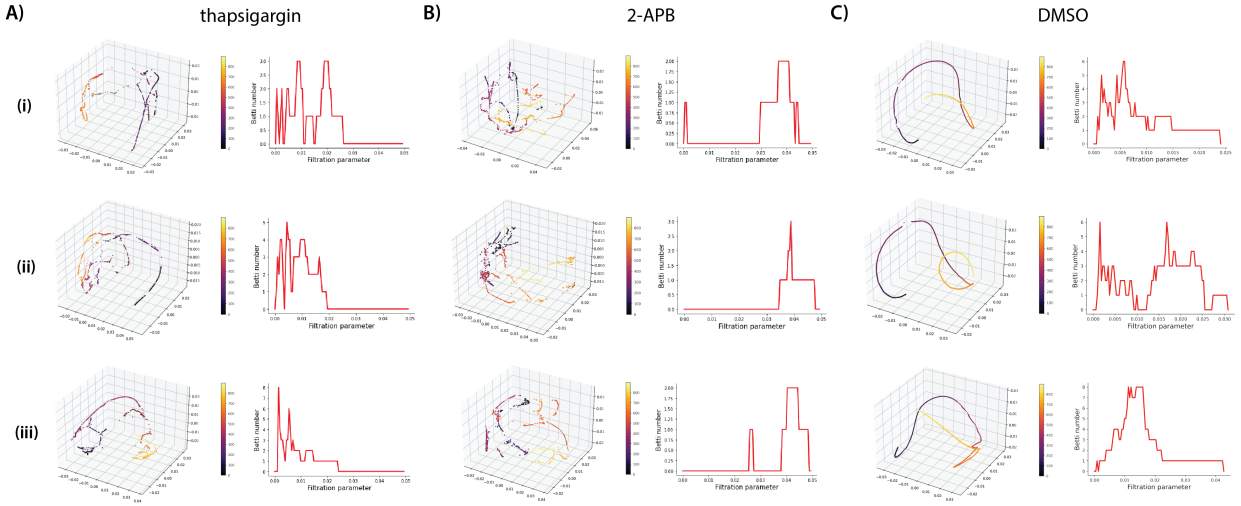

Figure 3: **GSTH analysis of various drug treatments in epidermal signaling.** PHATE visualization and Betti curves of dimension 1 homology computed using GSTH from the cell adjacency graph of the epidermal Ca^{2+} signaling dataset.

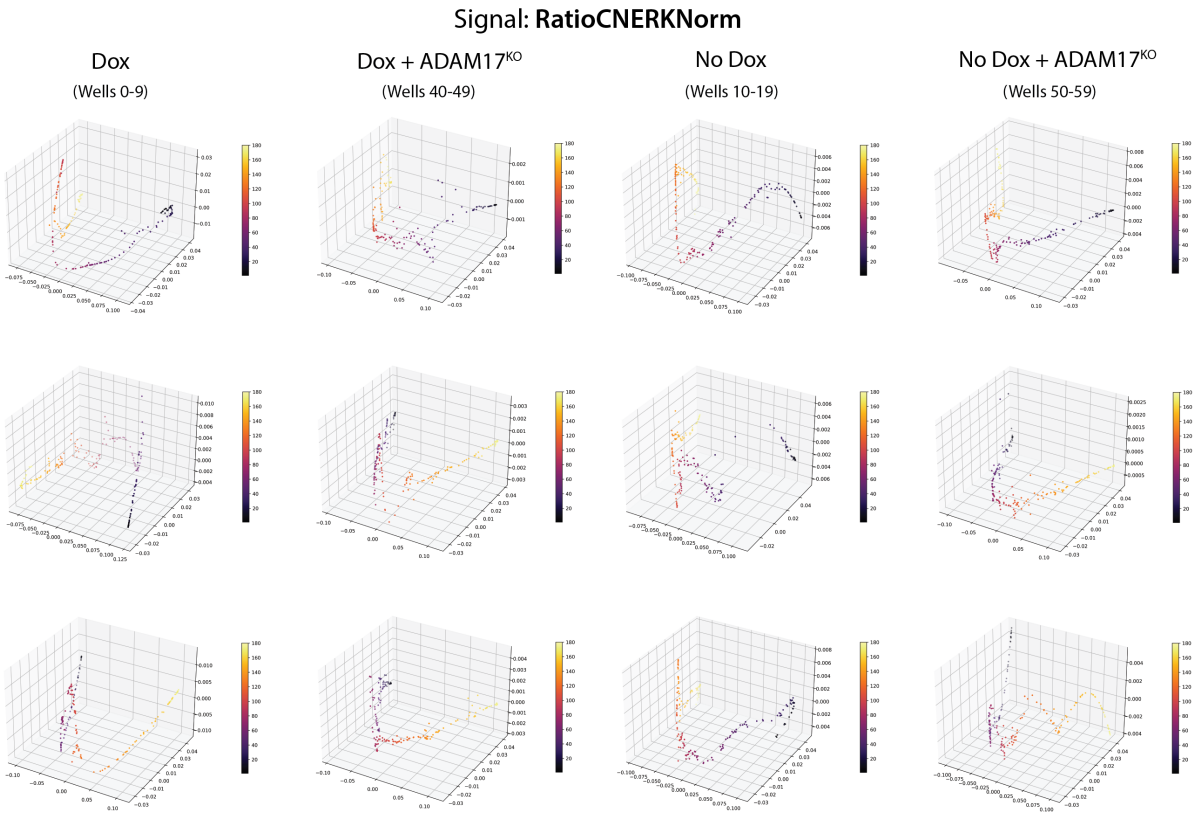

Figure 4: **GSTH analysis of various experimental conditions in the ERK signaling dataset.** PHATE visualization and Betti curves of dimension 1 homology computed using GSTH from the cell adjacency graph of the epidermal Ca^{2+} signaling dataset.

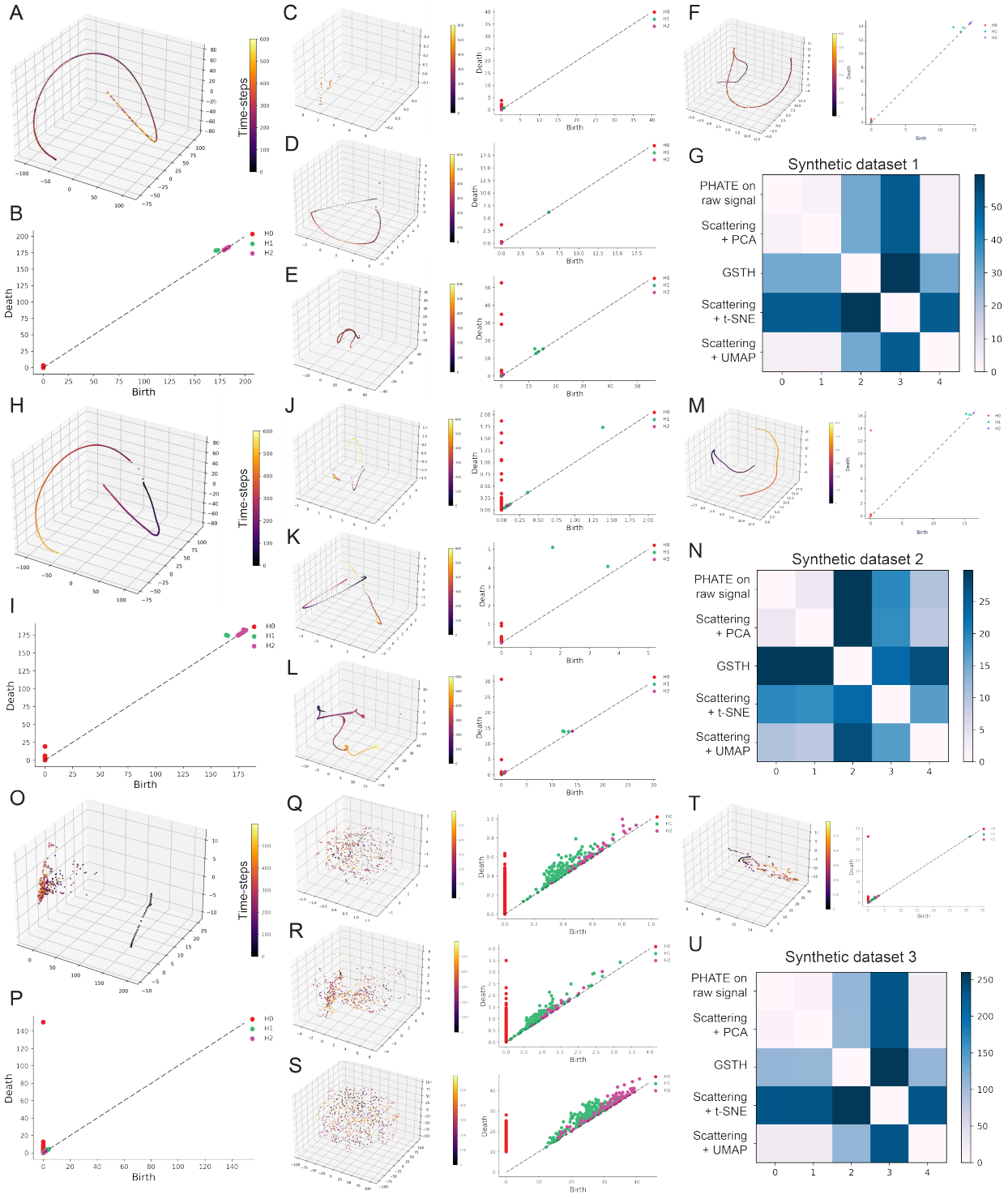

Figure 5: **Validation of GSTH using synthetic datasets.** PHATE visualization **(A, H, O)** and persistence homology **(B, I, P)** on synthetic data using GSTH and comparison with (1) directly applying PHATE on the input signals **(C, J, Q)**; (2) PCA on generated scattering coefficients **(D, K, R)**; (3) t-SNE on generated scattering coefficients **(E, L, S)**; (4) UMAP on generated scattering coefficients **(F, M, T)**. Finally, Wasserstein distances from the persistence diagrams of each methodology for each of the 3 synthetic datasets **(G, N, and U)**